A METHOD AND SYSTEM FOR UPDATING A SEARCHABLE DATABASE OF DESCRIPTIVE INFORMATION DESCRIBING INFORMATION STORED AT A PLURALITY OF ADDRESSABLE LOGICAL LOCATIONS

FIELD OF THE INVENTION

The invention generally relates to providing a group of users with a search facility for information stored at a plurality of addressable logical locations. Particular embodiments of the invention relate to a method and system for updating a database searchable by a group of users and containing descriptive information and addresses for the stored information.

BACKGROUND OF THE INVENTION

The provision of a search capability for information stored at a plurality of addressable locations is a problem when the amount of information becomes large and distributed. It is known in the art to provide a database of index information which is searchable to enable the address of the stored information to be located based on descriptive information stored with the address in the database.

With the prevalent use of the Internet, and in particular the World Wide Web, the problem of searching for and retrieving information in the form of web pages has received much attention. Many search engines have been developed that search and catalog web pages to form a database of addresses and descriptive information for those addresses. A user is thus able to submit a query to the search engine to search the database and to retrieve web pages best matching the query.

The problem with many prior art search engines is that they try to cover the whole of the World Wide Web. This is an almost impossible task in view of the fluid nature of the Internet. Also, many of the results of the search will not be relevant to the user's interests. Further, the requirement for cataloging the whole of the Internet places a vast burden on the processing power required.

One prior art system disclosed in US 5931907 comprises the local storage of information as a distributed database by a community of agents. When a page is loaded and considered to be of interest to a user, the agent can be instructed to catalog the page and the user can add additional user information. Other users of agents within the community can be notified of the potentially interesting information. In this way a community of users have access to potentially interesting information distributed across the network.

One disadvantage of this arrangement is that the information is not held centrally at the database and requires each of the agents to communicate with each of the other agents within the network. Further, the cataloging of web pages is initiated manually after a user has inspected the page.

SUMMARY OF THE INVENTION

Embodiments of the invention provide an improved search facility for information such as web pages to a group of users with common interests.

In accordance with one embodiment of the invention, there is provided a method and system for providing a group of users having a common interests with a search facility for information stored in a plurality of addressable logical locations. A database of index information for information that is stored at the plurality of logical locations is provided in which the index information includes the addresses of the logical locations and descriptive information for information stored at each logical location. The descriptive information matches a common profile of interests of the group of users. The accessing and retrieval of stored information by a user in the group is monitored and descriptive information is derived using the retrieved information. The relevance of the retrieved information is determined by comparing the descriptive information to the profile, and if any relevant retrieved information is determined, the database is updated using the address and descriptive information of the retrieved information that is determined relevant.

Embodiments of the invention can be implemented in a single apparatus such as a suitably programmed general purpose computer or dedicated hardware. However, preferred embodiments are applicable to a network wherein the database is provided at a server and the accessing and retrieval of stored information, the monitoring of the accessing and retrieval, and the deriving of the descriptive information takes place at a client. The address and the derived descriptive information is sent to the server for updating of the database at the server.

The determination of the relevance of retrieved information can take place in the client or in the server. Preferably the determination takes place in the client in order to reduce the amount of information transmitted to the server and to distribute the processing load.

In one embodiment an initial request from a client to access the database at the server is sent and an agent is downloaded from the server to the client in response. The agent comprises an autonomous application which when installed and running on the client performs the monitoring, determining and sending processes. The agent thus uses the profile to identify relevant information to be used to update the database. The application can be implemented in a multitasking environment in the background.

In a preferred embodiment, the user of the client is warned that in order to use the search facility, i.e. to be able to access the database, the agent must be downloaded. Access to the database is denied if no agent is installed on the client. Only when the user inputs a confirmation that is sent by the client to the server is the agent downloaded to the client from the server.

Thus the trade-off by a user for access to the search facility is that their computer is used to monitor their activities to contribute towards updating the database. The user is a member of a group of users who have a common interest and thus the agent has a profile representative of the common interest of the group. Thus for the user to access the database, the user allows the distributive

processing of information locations visited in order to update the database for the common good of the group.

Embodiments of the invention are suited to any system in which information is stored at addressable logical locations, and are particularly suited to the Internet in which the Internet Protocol is used and the stored information includes hypertext mark-up language (HTML) files. The logical addresses thus comprise Uniform Resource Locators (URLs). In this embodiment the client implements a web browser to access web pages hosted by web servers and the agent on the client monitors the accessing and retrieval of web pages.

In addition to the monitoring of the pages actually visited by the client, the agent can also include a "spidering" capability. Links from the web pages accessed and retrieved can be "spidered" or crawled by the agent in order to access and retrieve the web pages and determine descriptive information for further expanding the updating of the database. The web pages that are spidered or crawled are processed to determine descriptive information and the descriptive information is then analyzed to determine the relevance of the page. In this way index information only for relevant pages is sent to the server.

In one embodiment, the database is periodically checked to see if there are any entries in the database that have not been recently updated. If there are any entries that have not recently been updated, the web page can be accessed and retrieved by a spidering function at the server. Descriptive information for the page can then be determined and compared with the profile to determine if the page is relevant still. If the page is not relevant it is deleted from the database. If the page is relevant the entry in the database is updated with the new descriptive information and a date to show when it was updated.

The profile can comprise any information suitable for defining the common interests of the groups of users. When the stored information includes text such as web pages, the profile comprises descriptive information which comprises text.

The determination of relevance can then be performed on a keyword basis by matching the keywords of the profile to keywords in the descriptive information. The keyword matching need not be exact and can be based on lexical matching of synonyms. As an alternative matching technique, natural language matching of the text of the profile and the text of the descriptive information can be used.

Embodiments of the invention can be implemented on a single apparatus or on a client apparatus and a server apparatus each comprising a suitably programmed general purpose computer. Thus the invention can be embodied using computer program code for controlling a general purpose computer. The computer program code can be provided to a general purpose computer on any suitable carrier medium such as a storage medium (e.g. floppy disk drive, CD ROM, magnetic tape or programmable memory device) or a signal (such as an electrical signal carried over a network such as the Internet).

DESCRIPTION OF THE DRAWINGS

A preferred embodiment of the invention will now be described with reference to the accompanying drawings, in which:

Figure 1 is a schematic diagram of a system in accordance with a preferred embodiment of the invention,

Figure 2 is a flow diagram illustrating the process of downloading the agent from the server to the client in the preferred embodiment of the invention,

Figure 3 is a flow diagram illustrating the process of determining and sending descriptive information from the client to the server in the preferred embodiment of the invention,

Figure 4 is a flow diagram illustrating the process in the server for updating the database using the received information from the agent in the preferred embodiment of the invention, and

Figure 5 is a flow diagram illustrating the process of periodically updating the database in accordance with the preferred embodiment of the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Figure 1 schematically illustrates a system in accordance with a preferred embodiment of the invention for implementation as a search facility for web pages over the Internet 50. Clients 60 and 70 are connected to the Internet 50 in order to access web pages at web servers 30 and 40. The clients 60 and 70 have respective web browsers 61 and 71 implemented therein for accessing and retrieving web pages from the web servers 30 and 40. The clients 60 and 70 also have respective agents 62 and 72 loaded therein that have been downloaded to them in order to monitor the activity of the respective web browsers 61 and 71. The agents are autonomous applications that run in the background in a multi-tasking environment on the clients 60 and 70 such as in the WINDOWS operating system. The agents 62 and 72 are able to communicate through the Internet 50 to a search server 10 in order to communicate the results of their monitoring activities.

The search server 10 is connected to the Internet 50 to enable the clients 60 and 70 to access a search engine 3 via a web server 1 acting as the interface to the clients 60 and 70 using the web browsers 61 and 71. The search engine 3 interfaces to a database 20 providing a database of index information comprising logical addresses and descriptive information. In this embodiment the logical addresses comprise the URLs of web pages and the descriptive information comprises key words taken from the text of the web page (which can include the metatags). Thus a client 60 or 70 accesses the search server 10 using respective web browser 61 or 71 to access the web server 1 which acts as the interface and communicates with the search engine 3 to search the database 20. Thus the web server 1 and search engine 3 can provide a conventional search facility for searching database 20. However, the interface to the database 20 differs from conventional search engine interfaces in that when an initial request from a web browser 61 or 71 is received an agent download application 2 will detect whether the request comes from a client 60 or 70 which has an agent 62 or 72 loaded

thereon. If not the agent download application 2 will cause the web server 1 to warn the user of the web browser 61 or 71 that an agent must be downloaded in order to access the database 20 using the search engine 3. If the user inputs an acceptance, this is received from the web browser 61 or 71 by the web server 1 and passed to the agent download application 2 which downloads the agent 62 or 72 to the client 60 or 70. Thus when the web browser 61 or 71 next requests access to the search engine 3, access is permitted.

The search server 10 is also provided with a spider application 3 for carrying out the conventional spidering operation in order to periodically update the database 20.

The method of operation of the preferred embodiment of the invention will now be described in more detail with reference to the flow diagrams of Figures 2 to 5.

Figure 2 is a flow diagram illustrating the process of downloading the agent to the client in order to allow access to the search engine. When the client initially attempts to connect to the web server, in step S1 the browser is opened and used to access and retrieve a web page at the search server (step S2). The request to retrieve the web page from the browser will also include a request to access the search engine. In step S4 the search server detects whether there is an agent at the client. If an agent is present, in step S4 the client is allowed access to the search engine in order to search for web pages using keywords etc. in a conventional manner.

If the agent is not detected at the client (step S4), in step S6 a message is sent to the client and displayed to inform the user of the client that the agent must be downloaded in order to use the search engine. This message can be in the form of a web page with a check box to enable the user to accept the downloading of the agent in return for access to the search engine. In step S7 a user acceptance is then awaited. If no user acceptance is input, in step S8 the user is refused

access to the search engine. For example, if a user selects to decline downloading of the agent, a web page can be set to the web browser to inform the user that access to the search engine is refused.

Once the search server receives the acceptance from the agent, in step S9 the agent download application 2 downloads the agent to the client. The agent comprises an autonomous application capable of running in the background. The agent will include in the code or as metadata a profile defining the common interests of the group of users.

The profile can comprise a set of keywords. Once the agent has been downloaded in step S9, in step S10 the agent is installed from the client as is conventional in the WINDOWS operating system, in step S11 when the client is restarted, the agent runs automatically in the background. From then on the client is allowed access to the search engine (step S5). The installation of the agent on the client causes an icon to be added to the task bar in the WINDOWS operating system display. Thus the next time a user wishes to access the search engine, they can either use the web browser (step S1) or they can click on the agent icon in the task bar (step S3). If the agent icon is clicked on in the task bar, the web browser is launched and directed to access the search server. Alternatively the agent can include a web browser interface to act as the search interface for the client to access the search server to perform a search through the database 20.

The operation of the agent on the client will now be described in more detail with reference to the flow diagram of Figure 3.

In step S20 the client loads a web page from a web server 30 or 40. The agent picks up the URL and determines a catalog for the URL (step S1). The catalog can comprise any descriptive information. In this embodiment the process comprises the extraction of keywords from the hypertext mark up language (HTML) file. Methods for determining a catalog for a web page are well known in the art

and it will be apparent to a skilled person in the art that any known technique can be used for determining the catalog.

In step S22 the agent checks the catalog for the relevance of the page against a profile comprising key words that represent the interest of the group of users. Thus if in step S23 it is determined that the page is not relevant since the keywords for the page do not significantly match the key words for the profile, the page is ignored in step S24. If it is determined that the keywords (or a significant number of them) match, in step S25 the agent uploads the URL and the catalog to the search server. Once the URL and catalog have been uploaded to the search server if the page is relevant, in step S26 the agent determines whether the links on the page are to be cataloged. This can either be a preset parameter for the agent or the agent can determine this based upon the bandwidth (i.e. modem speed or LAN connection - or even mobile link speed) and processing power of the client. Also the server response time can be taken into account. If the links are not to be cataloged the process terminates in step S27. If the links are to be cataloged in step S28 the agent will determine the level of links to be cataloged. Once again the level of the links can be a predetermined number of links, or it can be based upon the processing power or bandwidth available to the client. This avoids too large a proportion of the processing power or communication bandwidth being taken up by the cataloging process (a spidering process) and avoids a significant downgrading of the performance of the users machine due to the "spidering" process. In step S29 it is determined whether the current cataloging level has been reached and if so the process terminates in step S27. If not in step S30 the agent searches for linked web pages which have not yet been cataloged. If there are noe in step S31 the process is terminated at step S27. If there are still linked web pages to be cataloged, in step S32 the agent sends a request and receives a linked page. The agent then determines a catalog for the URL in step S33 and the process returns to step S22 for the determination of the relevance of the page against the keywords.

It can thus be seen that the process of Figure 3 will continue until all of the pages to a predefined level have been cataloged and, where relevant the URLs and catalogs for the pages have been uploaded to the search server.

Thus in this embodiment of the invention not only is the page which has been visited by the web browser cataloged and used to update the database, also linked pages can be used to update the database. Thus the activity of the client machine is automatically monitored and when any relevant pages are detected these are used to update the central database for the good of the group of users. This ensures that when many clients are operating, the database is updated within the focus defined by the profile used by each of the agents. The profile in this embodiment comprises a carefully chosen selection of keywords.

The operation of the server upon receipt of the URL and catalog from the agent will now be described in more detail with reference to the flow diagram of Figure 4.

In step S40 a URL and catalog are received from the agent. It is then determined whether the URL is already in the database (step S41) and if so in step S42 it is determined whether the entry in the database has been updated recently or not. If it has been updated recently and the entry is not old (step S43) the URL and catalog are ignored and the process terminates in step S44. If the entry in the database for the URL is older than the predetermined age, in step S45 the received URL and catalog are used to update the URL and catalog in the database and the process proceeds to step S47.

If in step S41 it is determined that the URL is not in the database, in step S46 the URL and catalog received from the agent are added to the database. Once the database has either been added to or updated (step S46 or step S45) in step S47 the database entry is marked with the date so that the age of the entries in the database can be monitored particularly with regard to step S42.

In order to further expand the database, in step S48 the spider application within the search server requests and receives the page for the URL which has been added to or updated in the database. In step S49 the spider application then searches for any linked web pages on the received page. If there are none (step S50), the process terminates in step S44. If there are linked web pages, in step S51 it is determined whether the URLs are in the database. If so in step S52 it is determined whether the entries in the database are older than a predetermined age and if not the process terminates in step S44. If the entries are old (step S52) or if the URLs are not in the database, in step S53 the spider application requests and receives pages for the URLs. The spider application then determines catalogs for the pages in step S54 and in step S55 it is determined whether the pages are relevant or not by comparing the keywords in the catalog to the keywords stored as the profile. If the pages are determined not to be relevant, in step S56 the pages are ignored and in step S44 the process terminates.

If the pages are determined to be relevant (step S55) in step 57 the URLs and catalogs are added to or updated in the database. The database entries are then marked with the date in step S58 and the process terminates in step S44.

Thus in this process illustrated in Figure 4, the database is update using a catalog for the URL visited by the user of a client, catalogs for pages linked from the visited page as determined by the client, and catalogs for links from the visited web page as determined by the server.

The benefit of also providing for a spidering capability at the server is that the client may be provided with a limit spidering capability e.g., the level of the links to be followed by the spider in the client can be limited. This limits the processing power and bandwidth taken up by the agent. The full spidering process can thus be completed or indeed fully carried out by the server.

In addition to the spidering process carried out to supplement the catalogs received from the agents, the server can also periodically update the database.

This process will be described in more detail with reference to the flow diagram of Figure 5.

In step S60 periodically the spider application looks at the URLs in the database and in step S61 a determination is made as to whether any have not recently been updated. If all of the entries have recently been updated, the process terminates in step S68. If there are entries in the database which have not been recently been updated (step S61) the spider application requests and receives web pages for the URLs (step S62). The spider application then determines catalogs for the pages (step S63) and checks the relevance of the catalog against the keywords (step S64). If the pages are not relevant (step S65) in step S69 the URLs are deleted from the database and the process terminates in step S68.

If the pages are determined to be relevant (step S65) in step S66, the URLs and the catalogs in the database are updated and in step S67 the database entries are marked with the date of update. The process then terminates in step S68.

The process of Figure 5 thus comprises a conventional periodic spidering process in order to keep the database up to date. It enables the database to be pruned to remove pages that are no longer relevant.

Although a preferred embodiment of the invention has been described hereinabove, it will be apparent to a skilled person in the art that modifications lie within the spirit and scope of the invention.

For example although in the preferred embodiment the spider application 3 is illustrated as residing in the search server 10, the spider application can in fact reside on any physical server on the Internet 50. The spider application may then independently receive the URLs which are also sent to the search server for updating the database so that the spider can spider from these URLs. The resultant relevant links can then be submitted to the search engine much in the same way as relevant links are submitted by agents.

Although in the preferred embodiment the determination of the relevance of the page is implemented by the agent, alternatively, this function may be given to the search server. Thus in this case the agents 62 and 72 transmit catalogs and URLs for all pages visited by the web browser. Also catalogs and URLs for all links from the visited page can be sent to the search server. It can thus be left for the search server to determine the relevance of the pages for the updating of the database. This process is however less preferred since it increases the amount of data that has be to transmitted by the agents to the search server. Although in the preferred embodiment the matching process between the profile and the descriptive information (the catalog) was performed using keywords, alternative embodiments of the invention can be applied to the use of any form of descriptive information. The preferred embodiment of the invention is particularly suited to the use of text which can allow keyword matching either strictly or on the basis of synonyms or natural language matching of text. It is also possible to define a profile as comprising meta information such as the date of downloading into the web page by the web browser or the address of the originating site. The profile can comprise any information that allows for the definition of the common interests of the group of users using the clients 60 and 70.

Although in the preferred embodiment of the invention the network on which the clients and the search server are connected is described as comprising the Internet, further embodiments of the invention are applicable to any network and can for example comprise an Intranet, Extranet, or local area network. Further embodiments of the invention are more widely applicable to any form of information retrieval such as document retrieval over a network wherein a central database of index information is stored to allow for searching for a stored information.

The determination of the relevance of the stored information need not be based solely on the profile. The relevance can also be determined based on whether the database has recently been updated for that address.

In addition to updating the database using retrieved information which matches the profile, a user can select to update the database using any retrieved information by manually selecting it.

Further embodiments of the invention are not limited to the use of the Internet using web addresses, but may also be applicable to any logical addressing system and for example covers all protocols using URLs e.g. HTTP, FTP, POP, and SMTP.

Embodiments of the invention are ideally suited to the searching needs of a specific interest or community. The central database can self-focus, expand and update automatically based on the behavior of the members of the group. The common interests of the group can be defined by a suitable profile such as keywords and this keeps the domain of the search focused. However, the focusing of the search database does not prevent it being amended and expanded when users view a site that is not currently indexed. So long as the site falls within the current field of interest as defined by the profile, the site will automatically be indexed by the agent and the database updated.

Advantages of this arrangement are that the user community can focus on the development and usefulness of the search indexed over time. The users can update the search catalog database automatically themselves thus effectively distributing the processing task and requirement for bandwidth over many users.

In the preferred embodiment, the database is described as being updated as soon as a URL is passed from an agent, however, it is possible for the updating process to be modified such that the database is only updated when the URL is submitted by agents a predetermined number of times. This would indicate that one or a number of users visited the sight more than once, clearly indicating that the sight is relevant and should be added to the database.

It will be appreciated by those of ordinary skill in the art that the clients and servers described above may be implemented on computing devices controlled by appropriate programming instructions. Accordingly, embodiments of the invention may comprise a computing device including a processor to execute programming instructions and a storage device coupled to the processor and containing programming instructions for instructing the processor to perform data processing in accordance with various aspects of the invention. Appropriate storage devices may include but are not limited to volatile memory such as RAM, and non-volatile memory such as ROM or flash memory, and peripheral storage devices such as hard disks and optical disks.

The foregoing description relates to preferred embodiments of the invention. However, those having ordinary skill in the art will recognize a variety of alternative organizations and implementations that fall within the spirit and scope of the invention as defined by the following claims.